CEE 498DS: Data Science for CEE, Fall 2020

THIS SYLLABUS IS SUBJECT TO CHANGE! Please check back throughout the course.

Basic Course Information

- Department: Civil and Environmental Engineering
- Title: CEE 498DS: Data Science
- Credits: 3 for Undergraduates, 4 for Graduate Students
- Semester: Fall 2020
- Meeting time and location: 12-1:20 on Tuesdays and Thursdays on Zoom
- First day of instruction: 8/25/2020
- Last day of instruction: 12/8/2020

Basic Instructor Information

- Instructor: Prof. Christopher Tessum, PhD
- Office: 3213 Newmark Civil Engineering Laboratory
- Office hours: 10:30 am to 12:00 pm on Mondays and Wednesdays on Zoom
- Email: ctessum@illinois.edu
- Website: https://cee.illinois.edu/directory/profile/ctessum
- Names and contact information for teaching assistants: TBD

Description of the course

Welcome to CEE Data Science! This semester, you will learn to leverage data to study civil and environmental engineering problems, identify patterns, and make actionable insights. This course combines training in digital and computer tools—including distributed computing, exploratory data analysis, and statistical modeling and deep learning—with application of those tools to civil and environmental engineering issues.

This course differs from other available machine learning and data science courses in that it focuses on civil and environmental engineering problems and the methods used to solve them. In particular, this course emphasizes working with spatial data, which is common in physical science but less common in data science when applied to other disciplines.

By the end of the semester, you will be able to:

1. Use software tools for data processing and visualization, machine learning, and deep learning to
2. Retrieve, manipulate, and analyze data; and
3. Make inferences and predictions about the (built) environment.

This course will help you to gain the skills and tools necessary to make the most of the great increases in the amount and quality of data related to civil and environmental engineering that is being collected and stored.

Because data science methods are used across a number of different industries and instructional materials are readily available, this course will include readings and video lectures from across the internet. We will focus our face-to-face time on learning aspects of civil and environmental data science that differ from data science as used by other fields, and on applying data science concepts to solving physical problems. This course will be structured around semester-long projects; students will

choose project topics at the beginning of the semester and will apply the concepts learned in the class to their projects as the semester progresses.

Prerequisites

- CEE 202;
- CEE300, 330 or 360; and
- CS 101 or equivalent.

Course Structure

This course is structured as a series of modules, with each module containing recorded lectures, readings, and quizzes to be completed before each class meeting. Class meetings will be held on Zoom to go into further depth on the material that was covered in recorded lectures and readings. Near the beginning or the semester, students will choose a topic for a project, which they will work on throughout the semester, applying the concepts that we learn in class. Additionally, students will complete homework assignments and a midterm and final exam.

Course Requirements and Assessment Overview

- Grades will be assigned based on several types of deliverables:
  - Discussions: 3% of final grade
  - Homeworks: 17% of final grade
  - Midterm exam: 15% of final grade
  - Final exam: 25% of final grade
  - Course project: 40% of final grade: 5% for literature review, 15% for Kaggle competition, 2.5% for exploratory data analysis, 7.5% for final presentation, and 10% for final report.
- Graduate students are expected to register for 4 credits and undergraduates are expected to register for 3 credits. Correspondingly, course projects for graduate students are expected to include a machine learning component that is more complex than linear regression, whereas for undergraduates this is optional.
- Letter grades will be assigned according to the following scale:
  - 97-100: A+
  - 94-96.5: A
  - 90-93.5: A-
  - 87-89.5: B+
  - 84-86.5: B
  - 80-83.5: B-
  - 77-79.5: C+
  - 74-76.5: C
  - 70-73.5: C-
  - 67-69.5: D+
  - 64-66.5: D
  - 60-63.5: D-
  - Below 59.5: F

Homeworks and Exams

Homeworks and Exams will be done through [PrairieLearn](). In assigning these types of homeworks and exams, I'm placing emphasis on **mastery**. The idea is to keep doing questions until you master the

underlying concept or method. Once you do, you should be able to answer these questions very quickly.

Before doing the assessments, you will need to enroll in this class in PrairieLearn. To do so, go to https://prairielearn.engr.illinois.edu/pl/enroll and click on "Add course" next to "CEE 498DS: Data Science for Civil and Environmental Engineering, Fall 2020".

Important: When you log in to PrairieLearn, choose "Log in with Illinois" rather than "Log in with Google" or "Log in with Microsoft". The UIUC login is the only one that will work.

Homeworks

For the homeworks, I try to encourage preparation for class before a module starts, so if you finish all of the questions completely before the first meeting time for the module, you will receive 110% of the available points. Questions finished between the first and second meeting times of the module receive 100%, and questions finished up to two weeks after the module ends can receive 80%.

Note that new homeworks are assigned most weeks, so if you don't stay ahead, it can be easy to fall behind.

Exams

Exams are also administered using PrairieLearn. For exams, partial credit isn't given, but you can try each problem more than once, with a decreasing number of points possible for each try.

Time commitment

The University guidelines for course credit hours are posted here. In summary, students in a 3-credit class are expected spend at least 6 hours per week, and students in a 4-credit class are expected to spend at least 8 hours per week, working outside of class times on readings, assigned lectures, assignments, projects, and test preparation.

It is my goal for this class to follow these guidelines so let me know if you think it does not (keeping in mind that the guidelines are for the *minimum* effort requirements).

Learning Resources

- Students are expected to bring have use of a laptop for class.
- There is no required textbook to purchase. Course material will draw from a number of sources across the internet.
- Some supplemental textbooks which students may find useful are:
  - Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython
  - Hands-On Machine Learning with Scikit-Learn & TensorFlow
  - https://www.w3resource.com/python-exercises/

This manuscript is a template (aka "rootstock") for Manubot, a tool for writing scholarly manuscripts. Use this template as a starting point for your manuscript.

The rest of this document is a full list of formatting elements/features supported by Manubot. Compare the input (`.md` files in the `/content` directory) to the output you see below.

Basic formatting

Bold text

Semi-bold text

<div align="center">Centered text</div>

<div align="right">Right-aligned text</div>

Italic text

Combined *italics and **bold***

~~Strikethrough~~

1. Ordered list item
2. Ordered list item
   a. Sub-item
   b. Sub-item
      i. Sub-sub-item
3. Ordered list item
   a. Sub-item

- List item
- List item
- List item

subscript: H_2O is a liquid

superscript: 2^{10} is 1024.

[unicode superscripts]^{0123456789}

[unicode subscripts]$_{0123456789}$

A long paragraph of text. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Putting each sentence on its own line has numerous benefits with regard to [editing] and [version control].

Line break without starting a new paragraph by putting
two spaces at end of line.

Document organization

Document section headings:

Heading 1

Heading 2

Heading 3

Heading 4

Heading 5

Heading 6

A heading centered on its own printed page

Horizontal rule:

`Heading 1`'s are recommended to be reserved for the title of the manuscript.

`Heading 2`'s are recommended for broad sections such as *Abstract*, *Methods*, *Conclusion*, etc.

`Heading 3`'s and `Heading 4`'s are recommended for sub-sections.

Links

Bare URL link: https://manubot.org

[Long link with lots of words and stuff and junk and bleep and blah and stuff and other stuff and more stuff yeah](#)

[Link with text](#)

[Link with hover text](#)

[Link by reference](#)

Citations

Citation by DOI [1].

Citation by PubMed Central ID [2].

Citation by PubMed ID [3].

Citation by Wikidata ID [4].

Citation by ISBN [5].

Citation by URL [6].

Citation by alias [7].

Multiple citations can be put inside the same set of brackets [1,5,7]. Manubot plugins provide easier, more convenient visualization of and navigation between citations [2,3,7,8].

Citation tags (i.e. aliases) can be defined in their own paragraphs using Markdown's reference link syntax:

Referencing figures, tables, equations

Figure [1](#)

Figure [2](#)

Quotes and code

> Quoted text

> Quoted block of text
>
> Two roads diverged in a wood, and I—
> I took the one less traveled by,
> And that has made all the difference.

Code `in the middle` of normal text, aka `inline code`.

Code block with Python syntax highlighting:

```python
from manubot.cite.doi import expand_short_doi

def test_expand_short_doi():
    doi = expand_short_doi("10/c3bp")
    # a string too long to fit within page:
    assert doi == "10.25313/2524-2695-2018-3-vliyanie-enhansera-copia-i-
        insulyatora-gypsy-na-sintez-ernk-modifikatsii-hromatina-i-
        svyazyvanie-insulyatornyh-belkov-vtransfetsirovannyh-geneticheskih-
        konstruktsiyah"
```

Code block with no syntax highlighting:

```
Exporting HTML manuscript
Exporting DOCX manuscript
Exporting PDF manuscript
```
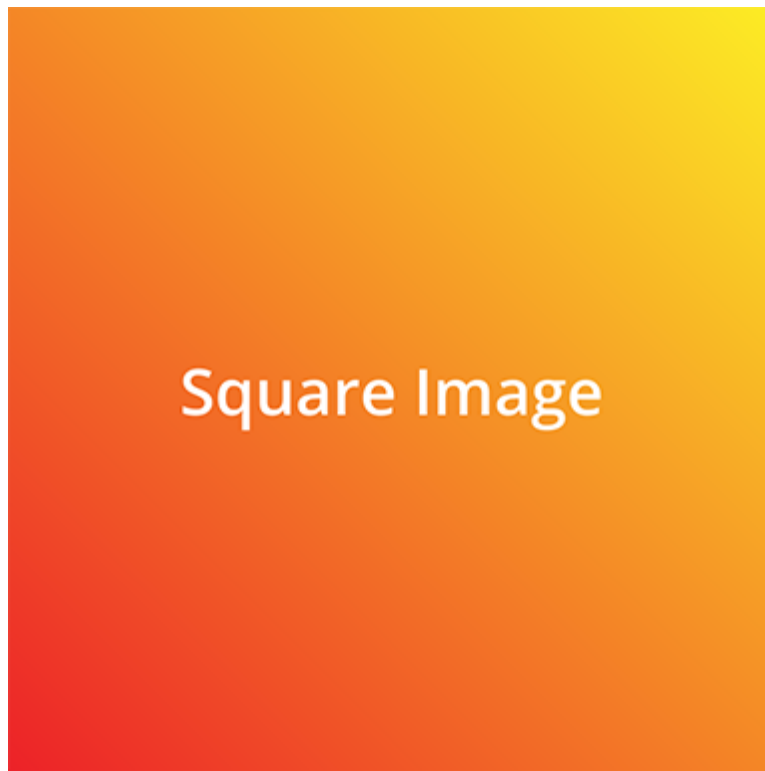
Figures

Figure 1:   A square image at actual size and with a bottom caption. Loaded from the latest version of image on GitHub.
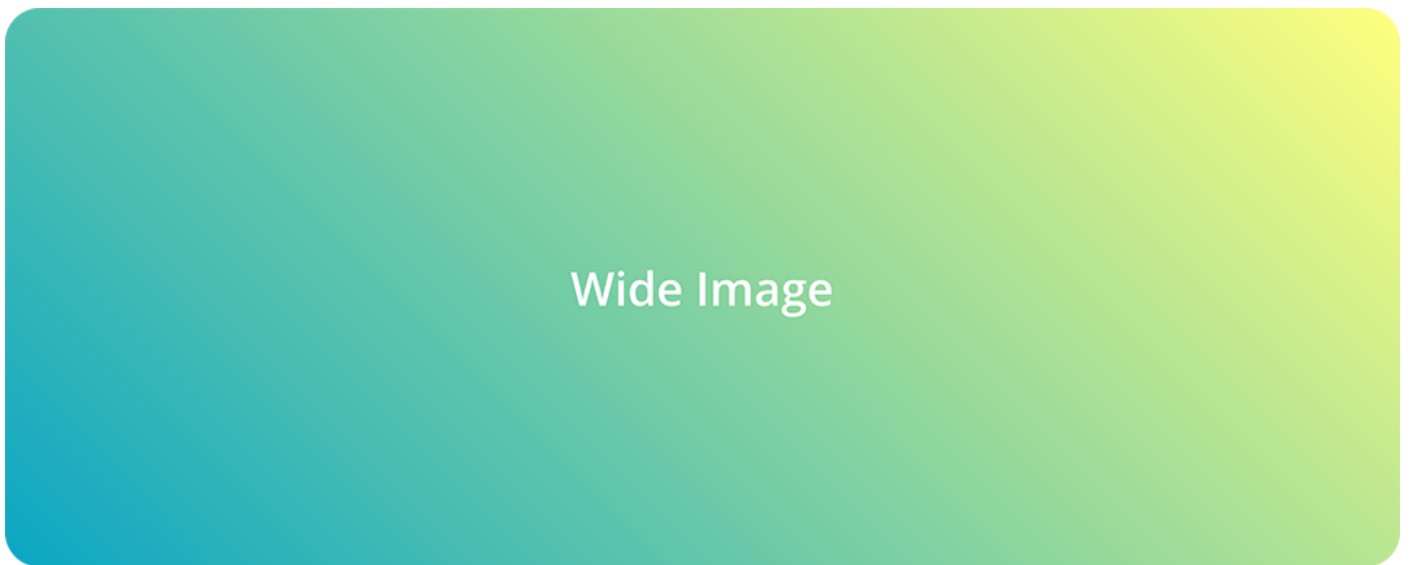

Figure 2:   An image too wide to fit within page at full size. Loaded from a specific (hashed) version of the image on GitHub.
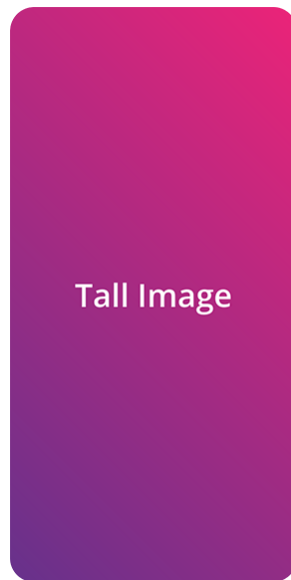
Figure 3: **A tall image with a specified height.** Loaded from a specific (hashed) version of the image on GitHub.

Figure 4: **A vector** `.svg` **image loaded from GitHub.** The parameter `sanitize=true` is necessary to properly load SVGs hosted via GitHub URLs. White background specified to serve as a backdrop for transparent sections of the image.

Tables

Table 1: A table with a top caption and specified relative column widths.

Bowling Scores	Jane	John	Alice	Bob
Game 1	150	187	210	105
Game 2	98	202	197	102
Game 3	123	180	238	134

Table 2: A table too wide to fit within page.

	Digits 1-33	Digits 34-66	Digits 67-99	Ref.
pi	3.14159265358979323 846264338327950	28841971693993751 0582097494459230	78164062862089986 2803482534211706	`piday.org`
e	2.71828182845904523 536028747135266	249775724709363999 5957496696762772	40766303535475945 7138217852516642	`nasa.gov`

Table 3: A table with merged cells using the `attributes` plugin.

	Colors	
Size	**Text Color**	**Background Color**
big	blue	orange
small	black	white

Equations

A LaTeX equation:

$$\int_0^\infty e^{-x^2}dx = \frac{\sqrt{\pi}}{2}$$

(1)

An equation too long to fit within page:

$$x = a + b + c + d + e + f + g + h + i + j + k + l + m + n + o + p + q + r + s + t$$
$$+ u + v + w + x + y + z + 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9$$

(2)

Special

⚠ **WARNING** *The following features are only supported and intended for* `.html` *and* `.pdf` *exports. Journals are not likely to support them, and they may not display correctly when converted to other formats such as* `.docx`*.*

    LINK STYLED AS A BUTTON

Adding arbitrary HTML attributes to an element using Pandoc's attribute syntax:

> Manubot Manubot Manubot Manubot Manubot. Manubot Manubot Manubot Manubot. Manubot Manubot Manubot. Manubot Manubot. Manubot.

Adding arbitrary HTML attributes to an element with the Manubot `attributes` plugin (more flexible than Pandoc's method in terms of which elements you can add attributes to):

> Manubot Manubot Manubot Manubot Manubot. Manubot Manubot Manubot Manubot. Manubot Manubot Manubot. Manubot Manubot. Manubot.

Available background colors for text, images, code, banners, etc:

`white` `lightgrey` `grey` `darkgrey` `black` `lightred` `lightyellow` `lightgreen` `lightblue` `lightpurple` `red` `orange` `yellow` `green` `blue` `purple`

Using the [Font Awesome](#) icon set:

✔ ? ★ 🔔 ⊗ ⋯

Policies

Inclusive Environment

The effectiveness of this course is dependent upon the creation of an encouraging and safe classroom environment. Exclusionary, offensive or harmful speech (such as racism, sexism, homophobia, transphobia, etc.) will not be tolerated and in some cases subject to University harassment procedures. We are all responsible for creating a positive and safe environment that allows all students equal respect and comfort. I expect each of you to help establish and maintain and environment where you and your peers can contribute without fear of ridicule or intolerant or offensive language.

If you witness or experience racism, discrimination, micro-aggressions, or other offensive behavior, you are encouraged to bring this to the attention of the course director if you feel comfortable. You can also report these behaviors to the Bias Assessment and Response Team (BART) (https://bart.illinois.edu/). Based on your report, BART members will follow up and reach out to students to make sure they have the support they need to be healthy and safe. If the reported behavior also violates university policy, staff in the Office for Student Conflict Resolution may respond as well and will take appropriate action.

Accommodations

To obtain disability-related academic adjustments and/or auxiliary aids, students should contact both the instructor and the Disability Resources and Educational Services (DRES) as soon as possible. You can contact DRES at 1207 S. Oak Street, Champaign, (217) 333-1970, or via email at disability@illinois.edu.

Participation

Active participation in the online learning environment is vital to your success in this course. Depending on your course, you may be asked to engage in online discussions and other interactive learning environments that invite your active participation and involvement with other students and your instructor.

Student Commitment

By registering for this online course, you commit to self-motivated study, participation in online course activities, and timely submission of all assignments. Furthermore, you commit to accessing the course

website and checking email at least four days per week (daily for 4-week courses), as well as to devoting at least 6–8 hrs./week (16-week course), 12–16 hrs./week (8-week course), or 24–32 hrs./week (4-week course) to preparing for each module and completing the required assignments and readings.

Deadlines

If you are unable to meet a particular deadline, it is your responsibility to make prior arrangements with the instructor for that given week. Otherwise, work submitted later than 1 day late will receive 10% penalty, and work submitted later than 2 days late will not be considered for grading unless consent has been given by the instructor.

Regrades

Requests for regrading homeworks, exams, and project deliverables must be submitted in writing within one week of receiving the initial grade.

Instructor Responses

Instructor Feedback Turnaround Time

Questions posted to the Course Help Discussion Forum generally will be answered within 48 hours. If possible, students are encouraged to answer questions posted by other students to the Course Help Discussion Forum, rather than waiting for an instructor's response.

Assignments submitted online will be reviewed and graded by the course instructor within 5 business days. Exams, essays, and term papers will be graded within 10 business days. If your instructor is unable to meet this timeline, students will be notified.

Contacting the instructor

For the fastest response response, the best way to contact the instructor is by attending office hours or posting questions to the Course Help Discussion Forum.

The instructor will not respond to phone calls. The instructor will respond to email messages within 48 hours of receiving them unless the instructor notifies you ahead of time of an inability to do so. When sending email, include a subject line that identifies the course number and nature of your question. The instructor may not respond to questions sent to him or her that should be posted in the Course Help Discussion Forum. Please don't be offended if you are asked to forward your question to this location.

Responding to the Discussion Forums

The role of the instructor within the discussion forums is to help facilitate discussion by providing probing questions, asking for clarification, and helping solve conflicts as necessary. The instructor will not respond to every post. You are encouraged to share your thoughts, experiences, and ideas with each other as well.

Academic Integrity

Academic dishonesty will not be tolerated. Examples of academic dishonesty include the following:

- Cheating

- Fabrication
- Facilitating infractions of academic integrity
- Plagiarism
- Bribes, favors, and threats
- Academic interference
- Examination by proxy
- Grade tampering
- Non-original works
- Should an incident arise in which a student is thought to have violated academic integrity, the student will be processed under the disciplinary policy set forth in the Illinois Academic Integrity Policy. If you do not understand relevant definitions of academic infractions, contact your instructor for an explanation within the first week of class.

Giving and receiving advice on projects and homework assignments is acceptable and encouraged. However, it is expected that help be given in general terms and in the form of natural language sentences (for example, English) rather than in the form of mathematical equations, algorithms, computer code, or anything else that could be copied and pasted into the recipient's answer. Similarly, students are encouraged to consult the Internet, but copying and pasting code from the Internet and submitting it for the class is not acceptable. The work that each student submits is expected to be their own, written with their own hand or typed on their own keyboard.

Copyright

Student Content

Participants in University of Illinois courses retain copyright of all assignments and posts they complete; however, all materials may be used for educational purposes within the given course. In group projects, only the portion of the work completed by a particular individual is copyrighted by that individual. The University of Illinois may request that students' materials be shared with future courses, but such sharing will only be done with the students' consent. The information that students submit during a course may, however, be used for the purposes of administrative data collection and research. No personal information is retained without the students' consent.

Non-student Content

Everything on this site and within University of Illinois courses is copyrighted. The copyrights of all non-student work are owned by the University of Illinois Board of Trustees, except in approved cases where the original creator retains copyright of the material. Copyrights to external links are owned by or are the responsibility of those external sites. Students are free to view and print material from this site so long as

- The material is used for informational purposes only.
- The material is used for noncommercial purposes only.
- Copies of any material include the respective copyright notice.
- These materials may not be mirrored or reproduced on non-University of Illinois websites without the express written permission of the University of Illinois Board of Trustees. To request permission, please contact the academic unit for the program.

Student Behavior

Student Conduct

Students are expected to behave in accordance with the penal and civil statutes of all applicable local, state, and federal governments, with the rules and regulations of the Board of Regents, and with university regulations and administrative rules.

For more information about the student code and handbook, see the CITL course policies page.

Netiquette

In any social interaction, certain rules of etiquette are expected and contribute to more enjoyable and productive communication. The following are tips for interacting online via email or discussion board messages, adapted from guidelines originally compiled by Chuq Von Rospach and Gene Spafford (1995):

- Remember that the person receiving your message is someone like you, deserving and appreciating courtesy and respect.
- Be brief; succinct, thoughtful messages have the greatest effect.
- Your messages reflect on you personally; take time to make sure that you are proud of their form and content.
- Use descriptive subject headings in your emails.
- Think about your audience and the relevance of your messages.
- Be careful when you use humor and sarcasm; absent the voice inflections and body language that aid face-to-face communication, internet messages are easy to misinterpret.
- When making follow-up comments, summarize the parts of the message to which you are responding.
- Avoid repeating what has already been said; needless repetition is ineffective communication.
- Cite appropriate references whenever using someone else's ideas, thoughts, or words.

Communications

Daily Contact

Your daily contact should be via the discussion forums in our Learning Management System and via email.

Course Questions

Questions pertaining to the course should be posted in our [Course Help Discussion Forum](#). You can get to this forum from the course home page. Posting questions here allows everyone to benefit from the answers. If you have a question, someone else is probably wondering the same thing. Anyone submitting a question via email will be directed to resubmit the question to the [Course Help Discussion Forum](#). Also, participants should not hesitate to answer questions posed by peers if they know the answers and the instructor has not yet responded. This not only expedites the process but also encourages peer interaction and support.

Personal and Grade-Related Questions

Questions of a personal nature should first be sent to the instructor's email address (listed on the Instructor Information page). When sending email, include a subject that identifies the course number and nature of your question.

Emergencies

If you have an emergency that will keep you from participating in the course, please notify your instructor by using the instructor's email address (listed on the Instructor Information page). Provide callback information in your email (if necessary). You should also notify your program director of any emergencies.

Zoom

Zoom is a tool that allows multiple people to join together simultaneously via a computer to text chat, audio chat, video chat, collaborate on a digital whiteboard, and even share their computer desktops with one another. The instructor's Virtual Office (when available) makes use of Zoom.

Instructor's Virtual Office

Another way to communicate with the instructor is to make use of the Virtual Office hours through the Zoom Interface. The instructor will be available for office hours via Zoom on the dates and during the times listed on the Virtual Office page in the Syllabus.

Announcements

The Announcements forum serves as a way for your instructor and University of Illinois administrators to make announcements within our online learning environment. Announcements posted here will also be sent to your Illinois email address, so be sure to check your email or the Announcements forum at least once a day to see whether any new announcements have been made.

Sexual Misconduct Policy and Reporting

The University of Illinois is committed to combating sexual misconduct. Faculty and staff members are required to report any instances of sexual misconduct to the university's Title IX and Disability Office. In turn, an individual with the Title IX and Disability Office will provide information about rights and options, including accommodations, support services, the campus disciplinary process, and law enforcement options.

A list of the designated university employees who, as counselors, confidential advisors, and medical professionals, do not have this reporting responsibility and can maintain confidentiality, can be found in the Confidential Resources section. Other information about resources and reporting is available at wecare.illinois.edu.

Student Wellness Resources

The University of Illinois strives to promote student success through the support of student psychological and emotional well-being. Please take advantage of the resources listed on the Student Affairs website.

Schedule

[Course calendar](#)

498 Data Science For CEE

Print Week **Month** Agenda ▼

Sun	Mon	Tue	Wed	Thu	Fri	Sat
Nov 1	2 Homework Assign / Project Activity As / **4:59am** 100% credi / **4:30pm** DS-CEE Of	3 **5pm** 110% credit H / **6pm** DS-CEE Zoom	4 **4:30pm** DS-CEE Of	5 **6pm** DS-CEE Zoom	6	7
8	9 Homework Assign / Project Activity As / **4:59am** 100% credi / +2 more	10 **5pm** 110% credit H / **6pm** DS-CEE Zoom	11 **4:30pm** DS-CEE Of	12 **6pm** DS-CEE Zoom	13	14
15	16 **4:59am** 100% credi / **4:59am** 80% credit / **4:30pm** DS-CEE Of	17 **5pm** 110% credit H / **6pm** DS-CEE Zoom	18 **4:30pm** DS-CEE Of	19 **6pm** DS-CEE Zoom	20	21
22	23 **4:59am** 100% credi / **4:59am** 80% credit / **4:59am** Project Acti / **4:30pm** DS-CEE Of	24 **6pm** DS-CEE Zoom	25 **4:30pm** DS-CEE Of	26 **6pm** DS-CEE Zoom	27	28
29	30 **4:59am** 80% credit / **4:59am** Project Acti / **4:30pm** DS-CEE Of	Dec 1 **6pm** DS-CEE Zoom	2 **4:30pm** DS-CEE Of	3 **6pm** DS-CEE Zoom	4	5

+ Google Calendar

Modules

Module	Title	Start Date
1	Open Reproducible Science	8/24/2020
2	Data science topics for Civil and Environmental Engineering	8/31/2020
3	Programming review	9/7/2020
4	Data	9/14/2020
5	Exploratory data analysis (EDA)	9/21/2020
6	Network analysis	9/28/2020
7	Spatial data	10/5/2020
8	Midterm	10/12/2020
9	Intro to machine learning	10/19/2020
10	Neural Networks	10/26/2020
11	Convolutional Neural Networks	11/2/2020
12	Neural Networks for Sequences	11/9/2020
13	Random forests	11/16/2020
14	Fall Break	11/23/2020

Module	Title	Start Date
15	Final project presentations	11/30/2020

Discussions

Title	Assigned	Initial Post Due	Response Posts Due
Open Reproducible Science	8/24/2020	Thu 8/27/2020, 12:00 CDT	Tue 9/1/2020, 12:00 CDT
Data science topics for Civil and Environmental Engineering	8/24/2020	Tue 9/1/2020, 12:00 CDT	Thu 9/3/2020, 12:00 CDT

Homeworks

Title	Assigned	Deadline for 110% Credit	Deadline for 100% Credit	Deadline for 80% Credit
Open Reproducible Science	8/24/2020	Thu 8/27/2020, 12:00 CDT	Sun 9/6/2020, 23:59 CDT	Sun 9/20/2020, 23:59 CDT
Programming review	8/31/2020	Tue 9/8/2020, 12:00 CDT	Sun 9/13/2020, 23:59 CDT	Sun 9/27/2020, 23:59 CDT
Data	9/7/2020	Tue 9/15/2020, 12:00 CDT	Sun 9/20/2020, 23:59 CDT	Sun 10/4/2020, 23:59 CDT
Exploratory data analysis (EDA)	9/14/2020	Tue 9/22/2020, 12:00 CDT	Sun 9/27/2020, 23:59 CDT	Sun 10/11/2020, 23:59 CDT
Network analysis	9/21/2020	Tue 9/29/2020, 12:00 CDT	Sun 10/4/2020, 23:59 CDT	Sun 10/18/2020, 23:59 CDT
Spatial data	9/28/2020	Tue 10/6/2020, 12:00 CDT	Sun 10/11/2020, 23:59 CDT	Sun 10/25/2020, 23:59 CDT
Intro to machine learning	10/12/2020	Tue 10/20/2020, 12:00 CDT	Sun 10/25/2020, 23:59 CDT	Sun 11/8/2020, 23:59 CST
Neural Networks	10/19/2020	Tue 10/27/2020, 12:00 CDT	Sun 11/1/2020, 23:59 CST	Sun 11/15/2020, 23:59 CST
Convolutional Neural Networks	10/26/2020	Tue 11/3/2020, 12:00 CST	Sun 11/8/2020, 23:59 CST	Sun 11/22/2020, 23:59 CST
Neural Networks for Sequences	11/2/2020	Tue 11/10/2020, 12:00 CST	Sun 11/15/2020, 23:59 CST	Sun 11/29/2020, 23:59 CST
Random forests	11/9/2020	Tue 11/17/2020, 12:00 CST	Sun 11/22/2020, 23:59 CST	Sun 12/6/2020, 23:59 CST

Project Assignments

Title	Assigned	Due
Project literature review	9/7/2020	Sun 9/27/2020, 23:59 CDT
Project exploratory data analysis	9/21/2020	Sun 10/25/2020, 23:59 CDT
Project Kaggle Competition	10/12/2020	Sun 11/22/2020, 23:59 CST
Project Group Final Presentation	11/2/2020	Sun 11/29/2020, 23:59 CST

Title	Assigned	Due
[Project Group Final Report](#)	11/9/2020	Sun 12/6/2020, 23:59 CST

Exams

- Midterm Exam: Thu 10/15/2020, 12:00 CDT—Fri 10/16/2020, 12:00 CDT
- Final Exam: Fri 12/11/2020, 13:30 CST—Fri 12/11/2020, 16:30 CST

Modules

Module 1: Open Reproducible Science

Module 1 Overview

This module covers tools and methods for ensuring your work is correct, understandable, and reproducible.

Module 1 Learning Objectives

By the end of this module, you should be able to:

- List the Bash commands for doing different computer operations
- Describe how to organize data files to improve transparency and reproducibility
- Define the syntax for the Markdown text formatting language
- List the Git and GitHub commands and operations for performing different operations

Module 1 Readings and Lectures

Develop your answers to the discussion questions below while completing the readings and lectures.

- [Introduction to Earth Data Science Chapter 1](#)
- [Introduction to Earth Data Science Chapter 2](#)
- [Introduction to Earth Data Science Chapter 3](#)
- [Andrej Karpathy: Software 2.0](#)
- [NOVA: What Makes Science True?](#)

Module 1 Discussion

This module includes a discussion section to help you understand by articulating how the module content could be useful in your professional life. Consider the following questions:

- What does it mean to practice open and reproducible science, and how could you apply it to your academic or professional life?
- Although the readings and NOVA video mainly refer to academic science, how could they be relevant to science practiced in industry?
- For the "Software 2.0" essay: What is the author talking about? Instead of trying to understand every detail in the essay (although by the end of the semester you should be able to understand a lot of it), focus on the main message: What is Software 2.0 and what are its implications for how science is carried out?

Log in to the [module discussion forum](#) and make one initial post and two responses. Refer to the [Discussion Forum Instructions and Rubric](#) for instructions how to compose posts to the discussion forum, and how they will be graded.

The initial post for this module are due by Thu 8/27/2020, 12:00 CDT and all response posts are due by Tue 9/1/2020, 12:00 CDT.

Module 1 Homework

The homework for Module 1 covers the required readings and lectures and is available [here](#). General information about homework assignments is [here](#).

The homework for this module is due by Thu 8/27/2020, 12:00 CDT for 110% credit, by Sun 9/6/2020, 23:59 CDT for 100% credit, and by Sun 9/20/2020, 23:59 CDT for 80% credit.

Module 1 Topics for Zoom Meetings

- Tue 8/25/2020, 12:00 CDT: Introduction, syllabus, and getting to know one another ([slides](#))
- Thu 8/27/2020, 12:00 CDT: Discussion of readings and technology check: installing Python and related tools ([notes](#))

Module 2: Data science topics for Civil and Environmental Engineering

Module 2 Overview

In this module we will learn the types of Civil and Environmental Engineering problems that data science and machine learning can help to answer, and begin to think about topics for course projects.

Module 2 Learning Objectives

By the end of this module, you should be able to:

- interpret how data science can be used in support of CEE
- formulate a data science problem statement
- describe a strategy for solving the problem

Module 2 Readings and Lectures

Develop your answers to the discussion questions below while completing the readings and lectures.

- [Deep Learning State of the Art (This is an introduction to what is currently possible with data science)](#)
- [Tackling Climate Change with Machine Learning](#)
- [Kaggle data science competitions](#)
- [Earth Engine Case Studies](#)
- [OpenAQ.org](#)
- [CACES air quality data](#)
- [EIA Energy Data](#)
- [UCI Machine Learning Datasets](#)
- [Dataset for defect detection in Concrete](#)
- [Best Datasets for Machine Learning Data Science](#)
- [VisualData Dataset Discovery](#)

Module 2 Discussion

This module includes a discussion section to help you understand by articulating how the module content could be useful in your professional life. Consider the following questions:

- A big part of this course will be a semester-long project, where you will use a dataset to answer a question relevant to Civil or Environmental Engineering. What are some ideas you have for course projects, how are they related to Civil or Environmental Engineering, why do you think they would be useful, and what dataset would they be based on?

Log in to the [module discussion forum](#) and make one initial post and two responses. Refer to the [Discussion Forum Instructions and Rubric](#) for instructions how to compose posts to the discussion forum, and how they will be graded.

The initial post for this module are due by Tue 9/1/2020, 12:00 CDT and all response posts are due by Thu 9/3/2020, 12:00 CDT.

Module 2 Topics for Zoom Meetings

- Tue 9/1/2020, 12:00 CDT: Project introduction, brainstorming, and proposals (slides) [https://docs.google.com/presentation/d/1MjCbv3tA5FBN5Pu2rU670q1lk-aeQ1qmYR2vg8ai7j8/edit?usp=sharing]
- Thu 9/3/2020, 12:00 CDT: Group discussion regarding project topics

Module 3: Programming review

Module 3 Overview

This course makes extensive use of the Python programming language. By brushing up on our Python skills now, we will make the rest of the course easier.

Module 3 Learning Objectives

By the end of this module, you should be able to:

- Express abstract concepts using Python syntax
- Solve mathematical problems using Python

Module 3 Readings and Lectures

- Complete the tutorials at https://learnpython.org, including those under 'Learn the Basics', 'Data Science Tutorials', and 'Advanced Tutorials'. These readings may not include all the information you need to complete the homework, which will allow you to practice researching concepts on the internet.

Module 3 Homework

The homework for Module 3 covers the required readings and lectures and is available [here](#). General information about homework assignments is [here](#).

The homework for this module is due by Tue 9/8/2020, 12:00 CDT for 110% credit, by Sun 9/13/2020, 23:59 CDT for 100% credit, and by Sun 9/27/2020, 23:59 CDT for 80% credit.

Module 3 Topics for Zoom Meetings

- Tue 9/8/2020, 12:00 CDT: Practice excercises and Q&A. [Practice excercises](#)
- Thu 9/10/2020, 12:00 CDT: Practice excercises and literature review intro. [Slides](#)

Module 3 Project Assignment

Project literature review

Module 4: Data

Module 4 Overview

Data relevant to Civil and Environmental Engineering comes in a number of different formats, including:

- Tabular data (csv, excel): Rows represent observations, columns represent properties
- Raster/image data ([geotiff](#), [NetCDF](#), png, jpg)
- graph data ([dot](#), [gtfs](#))

Module 4 Learning Objectives

By the end of this module, you should be able to:

- Specify information as tabular, raster, and graph data
- Use software tools to load and manipulate tabular, raster, and graph data

Module 4 Readings and Lectures

- [tabular data](#)
- [Pandas for tabular data](#)
- [graph data structure](#)
- [graph properties](#)
- [graph lecture notes](#)
- [image data](#)
- [pillow image processing](#)

Module 4 Homework

The homework for Module 4 covers the required readings and lectures and is available [here](#). General information about homework assignments is [here](#).

The homework for this module is due by Tue 9/15/2020, 12:00 CDT for 110% credit, by Sun 9/20/2020, 23:59 CDT for 100% credit, and by Sun 10/4/2020, 23:59 CDT for 80% credit.

Module 4 Topics for Zoom Meetings

- Tue 9/15/2020, 12:00 CDT: Group exercises [slides](#)
- Thu 9/17/2020, 12:00 CDT: [Guided exercise](#) and [group exercise](#)

Module 5: Exploratory data analysis (EDA)

Module 5 Overview

The first step in a data science project is getting a feel for the dataset you are working with. This is called Exploratory Data Analysis (EDA).

Module 5 Learning Objectives

By the end of this module, you should be able to:

- Calculate relevant statistical properties of an unfamiliar dataset
- Visualize an unfamiliar dataset and describe its relevant properties

Module 5 Readings and Lectures

- [Exploratory Data Analysis](#)
- [Exploratory Data Analysis in Pandas](#) watch time 8:00 to 20:30
- mlcourse.ai notebooks [1](#), [2.1](#) and [2.2](#)

Module 5 Homework

The homework for Module 5 covers the required readings and lectures and is available [here](#). General information about homework assignments is [here](#).

The homework for this module is due by Tue 9/22/2020, 12:00 CDT for 110% credit, by Sun 9/27/2020, 23:59 CDT for 100% credit, and by Sun 10/11/2020, 23:59 CDT for 80% credit.

Module 5 Topics for Zoom Meetings

- Tue 9/22/2020, 12:00 CDT: Practice project EDA [Notebook](#)
- Thu 9/24/2020, 12:00 CDT: Image processing [Notebook](#)

Module 5 Project Assignment

Project exploratory data analysis

Module 6: Network analysis

Module 6 Overview

In the previous module we learned how to do exploratory data analysis for tabular data. This week we will work on exploratory data analysis for graph (network) data.

Module 6 Learning Objectives

By the end of this module, you should be able to:

- Load and manipulate graph data using the `networkx` python library
- Perform basic statistical analysis of graph data

Module 6 Readings and Lectures

- [Game of Thrones: Network Analysis](#)
- [Networkx introduction](#)
- Networkx [betweenness centrality](#), [degree centrality](#), and [shortest path](#)

Module 6 Homework

The homework for Module 6 covers the required readings and lectures and is available [here](#). General information about homework assignments is [here](#).

The homework for this module is due by Tue 9/29/2020, 12:00 CDT for 110% credit, by Sun 10/4/2020, 23:59 CDT for 100% credit, and by Sun 10/18/2020, 23:59 CDT for 80% credit.

Module 6 Topics for Zoom Meetings

- Tue 9/29/2020, 12:00 CDT: Network analysis exercise ([Notebook](#))
- Thu 10/1/2020, 12:00 CDT: Midterm review

Module 7: Spatial data

Module 7 Overview

Many Civil and Environmental Engineering analyses—for example transportation networks or environmental data—have a spatial component. This week we will work with spatial data.

Module 7 Learning Objectives

By the end of this module, you should be able to:

- Analyze raster (gridded) data using the XArray library
- Analyze vector (point, line, or polygon) data using the GeoPandas library

Module 7 Readings and Lectures

- [Xarray tutorial](#) videos 1–8. If you want to follow along, the link to the dataset they use in the video is broken, but there is another copy [here](#)
- [Intro to Geopandas](#) videos 1, 3–14 (video 2 is installation, which you can do with anaconda)
- (In these and all video lectures, you can adjust the playback speed by clicking 'Settings' in the lower-right corner of the playback window.)

Module 7 Homework

The homework for Module 7 covers the required readings and lectures and is available [here](#). General information about homework assignments is [here](#).

The homework for this module is due by Tue 10/6/2020, 12:00 CDT for 110% credit, by Sun 10/11/2020, 23:59 CDT for 100% credit, and by Sun 10/25/2020, 23:59 CDT for 80% credit.

Module 7 Topics for Zoom Meetings

- Tue 10/6/2020, 12:00 CDT: Geopandas ([Notebook](#))
- Thu 10/8/2020, 12:00 CDT: Xarray ([Notebook](#))

Module 8: Midterm

Module 8 Learning Objectives

By the end of this module, you should be able to:

Access, characterize, and visualize the data for their projects

Module 8 Topics for Zoom Meetings

- Tue 10/13/2020, 12:00 CDT: Midterm course eval, project questions, visualizations
- Thu 10/15/2020, 12:00 CDT: Midterm Exam and project support

Module 8 Project Assignment

Project Kaggle Competition

Module 9: Intro to machine learning

Module 9 Overview

This week we will cover the basics of machine learning, including gradient descent, generalization, representation, and regularization.

Module 9 Learning Objectives

By the end of this module, you should be able to:

apply fundamental machine learning concepts

Module 9 Readings and Lectures

- [Google machine learning crash course](#), starting with 'Introduction to ML' and ending after 'Regularization: Sparsity'

Module 9 Homework

The homework for Module 9 covers the required readings and lectures and is available [here](#). General information about homework assignments is [here](#).

The homework for this module is due by Tue 10/20/2020, 12:00 CDT for 110% credit, by Sun 10/25/2020, 23:59 CDT for 100% credit, and by Sun 11/8/2020, 23:59 CST for 80% credit.

Module 9 Topics for Zoom Meetings

- Tue 10/20/2020, 12:00 CDT: Linear regression ([Notebook](#))
- Thu 10/22/2020, 12:00 CDT: Linear regression part 2 ([Notebook](#))

Module 10: Neural Networks

Module 10 Learning Objectives

By the end of this module, you should be able to:

- Train a neural network to make predictions about CEE datasets
- Identify and debug common problems with neural network training

Module 10 Readings and Lectures

- [Google machine learning crash course](), starting with 'Logistic Regression' and ending after 'Training Neural Nets'

Module 10 Homework

The homework for Module 10 covers the required readings and lectures and is available [here](). General information about homework assignments is [here]().

The homework for this module is due by Tue 10/27/2020, 12:00 CDT for 110% credit, by Sun 11/1/2020, 23:59 CST for 100% credit, and by Sun 11/15/2020, 23:59 CST for 80% credit.

Module 10 Topics for Zoom Meetings

- Tue 10/27/2020, 12:00 CDT: Neural Networks ([Notebook]())
- Thu 10/29/2020, 12:00 CDT: Regularization ([Notebook]())

Module 11: Convolutional Neural Networks

Module 11 Overview

This week we will learn about convolutional neural networks for computer vision. Training these neural networks from scratch is computationally- and data-intensive, but good results can be achieved with less time and data using a technique called transfer learning.

Module 11 Learning Objectives

By the end of this module, you should be able to:

Apply transfer learning to solve computer vision problems

Module 11 Readings and Lectures

- [Convolutional neural network explainer]()
- [Transfer learning video]()
- [Transfer learning notebook]()

Module 11 Homework

The homework for Module 11 covers the required readings and lectures and is available [here](). General information about homework assignments is [here]().

The homework for this module is due by Tue 11/3/2020, 12:00 CST for 110% credit, by Sun 11/8/2020, 23:59 CST for 100% credit, and by Sun 11/22/2020, 23:59 CST for 80% credit.

Module 11 Topics for Zoom Meetings

- Tue 11/3/2020, 12:00 CST: Remember to vote! ([class canceled]())
- Thu 11/5/2020, 12:00 CST: Convolutional Neural Networks ([Notebook]())

Module 11 Project Assignment

Project Group Final Presentation

Module 12: Neural Networks for Sequences

Module 12 Overview

Time series data are common in civil and environmental engineering, and machine learning can be used to make future predictions, for example for weather and pollution forecasts and predictions of mechanical failure.

Module 12 Learning Objectives

By the end of this module, you should be able to:

- Create recurrent neural networks for time series predictions
- Apply recurrent neural networks to answer questions related to CEE

Module 12 Readings and Lectures

- [An illustrated guide to recurrent neural networks](#)
- [Illustrated guide to LSTM's and GRU's: A step by step explanation](#)
- [Understanding LSTM networks](#)
- [Time series forecasting tutorial](#)

Module 12 Homework

The homework for Module 12 covers the required readings and lectures and is available [here](#). General information about homework assignments is [here](#).

The homework for this module is due by Tue 11/10/2020, 12:00 CST for 110% credit, by Sun 11/15/2020, 23:59 CST for 100% credit, and by Sun 11/29/2020, 23:59 CST for 80% credit.

Module 12 Topics for Zoom Meetings

- Tue 11/10/2020, 12:00 CST: LSTM for damage detection ([Notebook](#))
- Thu 11/12/2020, 12:00 CST: LSTM for damage detection ([Notebook](#))

Module 12 Project Assignment

Project Group Final Report

Module 13: Random forests

Module 13 Overview

In addition to linear regression and neural networks, random forest models are another commonly used machine learning framework

Module 13 Learning Objectives

By the end of this module, you should be able to:

Apply random forest models to answer questions related to CEE

Module 13 Readings and Lectures

- [Visual intro to decision trees part 1](#)
- [Visual intro to decision trees part 2](#)
- [Decision trees](#)
- [Regression trees](#)
- [Pruning regression trees](#)
- [Random forests part 1](#)
- [Random forests part 2](#)

Module 13 Homework

The homework for Module 13 covers the required readings and lectures and is available [here](#). General information about homework assignments is [here](#).

The homework for this module is due by Tue 11/17/2020, 12:00 CST for 110% credit, by Sun 11/22/2020, 23:59 CST for 100% credit, and by Sun 12/6/2020, 23:59 CST for 80% credit.

Module 13 Topics for Zoom Meetings

- Tue 11/17/2020, 12:00 CST:
- Thu 11/19/2020, 12:00 CST:

Module 14: Fall Break

Module 14 Overview

Happy Thanksgiving

Module 14 Topics for Zoom Meetings

- Tue 11/24/2020, 12:00 CST: No class
- Thu 11/26/2020, 12:00 CST: No class

Module 15: Final project presentations

Module 15 Learning Objectives

By the end of this module, you should be able to:

Students should have completed a project where they access and explore a civil or environmental dataset and use it to answer a scientific question.

Module 15 Topics for Zoom Meetings

- Tue 12/1/2020, 12:00 CST: Project presentations
- Thu 12/3/2020, 12:00 CST: Project presentations
- Tue 12/8/2020, 12:00 CST: Project presentations

Assessment Instructions and Rubrics

Discussion Forum Instructions and Rubric

This section describes how to participate in the discussion forum, and how your posts will be graded.

Initial Post

In the Discussion Forum for the module, compose an initial post that responds to at least one of the questions above. Your initial post is your opportunity to engage with the prompt in a way that is unique to you. Some ways to accomplish that include:

- Connect with the prompt in a personal way by incorporating personal anecdotes.
- Reflect on any potential biases you may have based on your experiences.
- Consider any potential biases in the information presented in the prompt itself. Be open to different points of view by providing some suggestions of what those might be.

Your initial post must meet the following requirements:

- Include at least **200 words**, excluding any references.
- Use appropriate evidence from the readings and lessons to support your claims and judgments.

Response Posts

Post at least 2 responses in the same thread. Your replies should stimulate more in-depth discussion about the topic. Some ways to accomplish that include:

- Clarify and/or extend your peers' line of thinking.
- Compare/contrast their views on the topic with your own.
- Suggest/question what explanation(s) you think your peers might be missing that could strengthen their arguments.
- End your response with a question to further the dialogue.

Your response posts should meet the following requirements:

- Include at least **50 words**, excluding references.
- Use of appropriate evidence from the readings and lessons to support your claims and judgments.

Submission Directions

- Access the discussion board and begin a new thread.
- Response Post: Select the title of any post to review it and read any replies already submitted. Click Reply next to any post to compose a reply.

Evaluation

This activity is worth 40 points: 20 points for your initial post and 10 points for each response post. Please see the rubric below for detailed information about how your posts will be graded.

Discussion Forum Rubric

Contributions	Description	Initial Post Points Assigned	Response Points Assigned

Contributions	Description	Initial Post Points Assigned	Response Points Assigned
Provocative	Response goes beyond simply answering the prompt; attempts to stimulate further thought & discussion	20	10
Substantial	Response provides most of the content required by the prompt, but does not require further analysis of the subject	15	7.5
Superficial	Response provides obvious information without further analysis of the concept; lacks depth of knowledge or reasoning	10	5
Incorrect	Response does not accurately address the prompt; rambling and/or without consistency	5	2.5
None	No response provided to the prompt within the associated timeframe	0	0

Course Project

This course includes a semester-long project, where you will apply the skills we learn in this class to apply data science to a problem relevent to Civil or Environmental Engineering. Project topics will be selected from a combination of student-proposed topics and topics selected by the instructor. The projects will include some individual components and some team components, with teams comprised of all the students assigned to each topic. The project will include a literature review, an exploratory data analysis midterm presentation, a kaggle competition, and a final report and presentation.

Project literature review

By the time this literature review is completed, students should be able to:

- Identify previous examples of work related to a given project,
- Describe what previous studies did, and
- Critique the previous studies, describing strengths and weaknesses.

Working as a team, students will identify prior work related to the project topic from sources including peer-reviewed literature and [ArXiv](#) (and other ArXiv-like repositories such as [EarthArXiv](#)). Students will then individually write an at least 400-word review of one of the articles. Instructions about how to write a literature review are available [here](#), for example.

Literature reviews should be submitted to the "course project" section on the compass2g course site.

Literature review rubric

	Exceptional (100%)	Satisfactory (85%)	Needs work (70%)	Incomplete (<60%)	
Description (10 pts)	Clearly describes the important aspects of the study methods and results	Description is clear but not complete, or complete but not clear	Description is missing key aspects of study methodology and results	Description missing majority of important aspects	
Critique (10 pts)	Clearly discusses the paper's strengths and weaknesses in meeting the goal of the student's project	Critique is clear but not complete, or complete but not clear	Critique misses important strengths and weaknesses	Critique does not accurately describe strengths and weaknesses	

Project exploratory data analysis

After completing this exploratory data analysis, students should be able to:

- Interpret a dataset in the context of using it to answer a question
- Select and implement statistical and graphical tools for describing a dataset
- Describe and defend reasons for choosing specific tools
- Analyze visual and statistical results to draw conclusions about the dataset

Each project will include a scientific question to be answered, a dataset relevant to the question, and a machine-learning method used to answer the question using the dataset. The second deliverable for the project is an expoloratory data analysis (EDA) of the dataset your project which uses visualizations and statistics to summarize the main characteristics of the dataset.

The EDA should be done by students individually as a notebook in the [Kaggle](#) data science platform. See [here](#) for an example EDA. The EDA notebook should read like a narrative, with natural language descriptions, code, statistics, and visualizations all interspersed with each other.

Each group member should perform and submit their own EDA. However, as detailed below, part of each student's score is based on the diversity of approaches taken by the different members of each project group, and part of each student's score is based on the average score of all members of the group. (Group members that do not submit the assignment will not be included in the overall average.) So it is in the best interest of group members to work together to ensure that a variety of different approaches are represented by the individual submissions, and to ensure that the individual submissions are all of high quality.

Exploratory data analysis rubric

	Exceptional (100%)	Satisfactory (85%)	Needs work (70%)	Incomplete (<60%)	
Technical content (20 pts)	Statistical and visual representations of the data thoroughly characterize all aspects of the data relevant to the project	Representations of the data characterize almost all relevant aspects of the data	Representations are present but not thorough	The dataset is not well described	
Graphical representation (10 pts)	Uses best practices for graphical display in all cases	Graphics mostly follow best practices, with some exceptions	Graphics are interpretable but do not follow best practices	Graphics are not interpretable	
Documentation (10 pts)	The document has a narrative flow, interspersing a well written introduction, method description including reasons for choosing the methods used, and discussion amidst the code and results	The document has a narrative flow but writing is not completely clear or some aspects are missing	Some documentation is included but it is not thorough	Documentation is not included	
Group diversity (1 pt)	Group submissions represent a diverse set of approaches	There is some diversity of approaches among group members	Group submissions are similar	Group submissions are very similar	
Group average (1 pt)	Average of group member scores == 100%	Average of group member scores == 85%	Average of group member scores == 70%	Average of group member scores < 60%	

Project Kaggle competition

After completing this competition, students should be able to:

- Choose a machine learning algorithm appropriate for a given dataset and question
- Implement and train the algorithm to a high degree of accuracy while avoiding overfitting
- Interpret the model results and performance

The class projects will be set up as Kaggle competitions, open only to members of this class. Working individually, they will choose an algorithm and use it to answer the class question. Students should extend the documentation narrative from their EDA to cover methods, results, and discussion for their learning algorithm. Here is an example of a combined EDA and model prediction. Only some of the project datset will be made available for training the model on; the rest will be held separately to test the model results. Students will be able to see how their model performance compares to that of the other students in their project team.

At the end of the project period, the student with the highest score in the competition gets 2 points extra credit on this project component. If there is a tie for first place, or if there are less than 4

students in the project group, no extra credit is awarded.

Each group member should submit their own entry to the competition. However, as detailed below, part of each student's score is based on the diversity of approaches taken by the different members of each project group, and part of each student's score is based on the average score of all members of the group. (Group members that do not submit the assignment will not be included in the overall average.) So it is in the best interest of group members to work together to ensure that a variety of different approaches are represented by the individual submissions, and to ensure that the individual submissions are all of high quality.

Kaggle competition rubric

	Exceptional (100%)	Satisfactory (85%)	Needs work (70%)	Incomplete (<60%)	
Algorithm (20 pts)	A best-in-class algorithm is chosen and implemented correctly	An appropriate alorithm is chosen and implemented correctly	An appropriate algorithm is chosen but not implemented correctly	An inappropriate algorithm is used	
Training (20 pts)	The algorithm is tuned to maximize predictive accuracy while avoiding overfitting	The algorithm is trained correctly but slightly underfits or overfits	There is substantial over- or under-fitting	The algorithm does not fit the data	
Documentation (10 pts)	The document has a narrative flow, interspersing a well written introduction, method description including reasons for choosing the methods used, and discussion amidst the code and results	The document has a narrative flow but writing is not completely clear or some aspects are missing	Some documentation is included but it is not thorough	Documentation is not included	
Group diversity (5 pts)	Group submissions represent a diverse set of approaches	There is some diversity of approaches among group members	Group submissions are similar	Group submissions are very similar	
Group average (5 pts)	Average of group member scores == 100%	Average of group member scores == 85%	Average of group member scores == 70%	Average of group member scores < 60%	

Project final presentation and report

After completing this presentation and report, students should be able to:

- Collaborate in a team to author a report and presentation
- Evaluate and defend their choice of machine learning methods

After the competition ends, students within each project team will collaborate to produce a final report and presentation that synthesizes the best aspects of all the individual analyses. Reports should include Introduction, Methods, Results, and Discussion sections, and should be long enough to thoroughly cover all of those topics. Because at this point the course projects are already implemented as narrative documents, they should be relatively straightforward to convert to a report.

Reports are **required** to be prepared using the Manubot manuscript preparation system, which is an alternative to MS Word. This may seem cruel and unusual, but there are two reasons:

1. Manubot makes extensive use of Markdown, Git, and Github, so using it reinforces the skills we are learning in class, and
2. Github has a mechanism for tracking contributions, which allows the instructor to grade project members according to their level of contribution.

The presentation should cover the material in the report and be 10–12 minutes in length. One person or multiple people can present. The presentation will be stopped at 12 minutes regardless of whether it's finished.

Final presentation and report rubric

	Exceptional (100%)	Satisfactory (85%)	Needs work (70%)	Incomplete (<60%)	
Writing (10 pts)	All required sections are present, well organized, and clearly written	There are minor defects in organization or clarity	There are substantial defects in organization or clarity	Not all required sections are present or interpretable	
Graphical representation (10 pts)	Uses best practices for graphical display in all cases	Graphics mostly follow best practices, with some exceptions	Graphics are interpretable but do not follow best practices	Graphics are not interpretable	
Oral presentation (5 pts)	Presentation is well organized, complete, and succinct	Minor issues with clarity, organization, or rambling	Poor organization or significantly too long	Presentation is not given	

Participation multiplier

The overall presentation and report grade will be multiplied by the following percent, determined by activity on github and in presentation:

- 1.0: Among the most active contributors to the project
- 0.85: Substantially lower contribution than the most active members
- 0.7: No apparent contribution
- 0.0: Statement from all team members that member did not contribute at all.

MS Word multiplier

Teams not using Manubot will have their presentation and report score multiplied by 0.9.

1. **Sci-Hub provides access to nearly all scholarly literature**
   Daniel S Himmelstein, Ariel Rodriguez Romero, Jacob G Levernier, Thomas Anthony Munro, Stephen Reid McLaughlin, Bastian Greshake Tzovaras, Casey S Greene

eLife (2018-03-01) https://doi.org/ckcj
DOI: 10.7554/elife.32822 · PMID: 29424689 · PMCID: PMC5832410

2. **Reproducibility of computational workflows is automated using continuous analysis**
   Brett K Beaulieu-Jones, Casey S Greene
   Nature biotechnology (2017-04) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6103790/
   DOI: 10.1038/nbt.3780 · PMID: 28288103 · PMCID: PMC6103790

3. **Bitcoin for the biological literature.**
   Douglas Heaven
   Nature (2019-02) https://www.ncbi.nlm.nih.gov/pubmed/30718888
   DOI: 10.1038/d41586-019-00447-9 · PMID: 30718888

4. **Plan S: Accelerating the transition to full and immediate Open Access to scientific publications**
   cOAlition S
   (2018-09-04) https://www.wikidata.org/wiki/Q56458321

5. **Open access**
   Peter Suber
   MIT Press (2012)
   ISBN: 9780262517638

6. **Open collaborative writing with Manubot**
   Daniel S. Himmelstein, Vincent Rubinetti, David R. Slochower, Dongbo Hu, Venkat S. Malladi, Casey S. Greene, Anthony Gitter
   Manubot (2020-05-25) https://greenelab.github.io/meta-review/

7. **Opportunities and obstacles for deep learning in biology and medicine**
   Travers Ching, Daniel S. Himmelstein, Brett K. Beaulieu-Jones, Alexandr A. Kalinin, Brian T. Do, Gregory P. Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M. Hoffman, … Casey S. Greene
   Journal of The Royal Society Interface (2018-04-04) https://doi.org/gddkhn
   DOI: 10.1098/rsif.2017.0387 · PMID: 29618526 · PMCID: PMC5938574

8. **Open collaborative writing with Manubot**
   Daniel S. Himmelstein, Vincent Rubinetti, David R. Slochower, Dongbo Hu, Venkat S. Malladi, Casey S. Greene, Anthony Gitter
   PLOS Computational Biology (2019-06-24) https://doi.org/c7np
   DOI: 10.1371/journal.pcbi.1007128 · PMID: 31233491 · PMCID: PMC6611653